融合句法信息的金融论坛文本情感计算研究*

兰秋军　刘文星　李卫康　胡星野
(湖南大学工商管理学院　长沙 410082)

摘要: 【目的】为了准确识别金融论坛文本的情感倾向，提出一种基于依存句法的情感分析方法。【方法】以依存句法的分析结果为基础，对句子进行情感主干抽取；然后根据依存关系的不同类型和不同的词性搭配，定义情感计算规则，以此进行句子情感倾向性计算。【结果】实验结果表明，该方法的整体准确率为 84.46%；看涨类的平均精确率和召回率分别为 82.84%和 87.14%，F 值为 84.94%；看跌类的平均精确率和召回率分别为 86.28%和 81.74%，F 值为 83.95%。【局限】在情感计算时未充分考虑子句间的关联关系。【结论】使用依存句法能有效提高金融论坛文本情感计算的准确性。

关键词: 情感分析　依存句法　金融论坛文本　文本挖掘
分类号: C931.6　G35

1 引　言

随着互联网的发展和普及，人们不再满足于被动地接受网络信息，越来越多的人开始在互联网上表达自己的观点和情绪。在这种背景下，文本情感分析技术应运而生。情感分析又称观点挖掘，属于自然语言处理范畴，旨在自动识别文本中人们对产品、服务、组织、事件等的评价、态度和情绪[1]。该技术无疑对了解大众情绪、把握舆情发展趋势、改善产品质量、提高服务水平等都具有非常巨大的潜在应用价值。而在金融领域，行为金融理论已表明，投资者的情绪是金融市场中的一个重要变量。以网络论坛、新闻、微博等为数据来源，应用情感分析技术挖掘市场中投资者的情绪，并以此作为投资决策依据的设想，已引发众多金融分析人士的关注[2-4]。然而，金融论坛语料具有短文本的相关特点，其特征稀疏、噪声大等特性给传统的情感分析方法带来了极大的挑战[5-6]。为了准确识别金融论坛文本的情感倾向，本研究以依存句法分析技术为基础，挖掘句子中各词语间的语义修饰关系，改善金融论坛语料情感分析的性能。

2 相关研究

目前针对金融论坛语料进行情感分析的技术主要有两类：基于情感词典的方法和基于机器学习的方法。其中，基于情感词典的方法最为简单，它主要依赖于词袋模型，将文本看成是一个无序的词汇集合，根据情感词典识别文本中的情感词，通过累加各词的情感分值，获得最终的文本情感倾向。如段江娇等[7]将情绪分为 5 个档次，根据帖子内容中的词汇与预先设定好的各档次关键词词库的匹配结果确定整个帖子的情绪。文献[8-10]探讨了投资者情绪对股票市场的影响，其采用的情感分析工具是武汉大学开发的 ROST 系统，该系统基于情感词典、程度词典等，获取文本中的情感词和程度词，进而判断单句的情感倾向。基于机器学习的方法是目前情感分析领域的主流，在对金融论坛语料的处理中更是常见。如文献[11-12]利用支持向

通讯作者：兰秋军，ORCID：0000-0001-7523-9487，E-mail：lanqiujun@hnu.edu.cn。

*本文系国家自然科学基金重点项目“高维度、非线性、非平稳及时变金融数据建模和应用”(项目编号：71431008)和国家自然科学基金面上项目“基于网络留言的投资者情绪测度模型、系统及应用”(项目编号：71171076)的研究成果之一。

量机的分类方法，对东方财富股吧、新浪股吧、和讯股吧等网络舆情信息进行情感分类，构建情绪指数并进行股票价格预测。文献[13-14]使用目前应用较为广泛的开源软件 Weka，并对比了多种算法，最终选择表现最好的 KNN 算法进行情感分类，同时构建情绪指数，研究其对股票市场的影响。

这两类方法中，情感词典方法的主要优势是思想和算法实现比较简单，它基于一个定义良好的情感词典，对各个词的情感分值进行简单累加。而机器学习方法无需情感词典，它能从大量语料中自动获取信息以构建情感计算模型，并在实际中有不错的表现。然而机器学习方法需要事先提供一个充分的、经过标注的语料库作为训练数据。必须指出的是，这两类方法目前均是以文本中的词语统计为基础，未对文本中深层的句法结构和语义关系进行分析和利用。

事实上，中文是一门非常复杂的语言，同样的词语在不同的句法结构下会产生不同的语义关系，进而形成迥然不同的情感色彩。因此，越来越多的学者开始使用句法分析来提高文本情感分析的准确性。如夏梦南等[15]在进行微博的情感分析时，利用句法分析和 CRFs 抽取候选评价对象，以此为基础使用 SVM 方法对微博进行情感分类。张庆庆等[16]通过依存句法解析，构造了由支配词、从属词、从属关系组成的三元组依存句法关系特征，并使用支持向量机和深度信念网络的方式对酒店评论语料进行情感分类。Nakagawa 等[17]将英语和日语的依存句法树作为 CRFs 模型的特征输入，对文本进行情感分类。肖红等[18]通过句法分析，获取词语在句子中扮演的不同角色(主、谓、宾、定、状、补)，对不同的角色给予不同的权值，以此计算句子的情感指数。上述方法从不同角度对语料中的句法信息加以应用，提升了情感分析的性能；然而却很少考虑到词语之间的修饰关系，尤其是修饰关系和词性的搭配对句子情感带来的影响。

本文借助依存句法分析技术，通过获取句子的句法结构和词语间的修饰关系进行情感传递，创新地将主谓宾关系和句子核心作为情感主干，并基于对大量语料的统计和观察分析，提出了若干情感计算规则，最终构建了情感计算模型。实验结果表明，该模型与传统的机器学习方法相比，在准确率和召回率上均有明显提升。

3 融合句法信息的情感计算

3.1 词典构建

根据情感分析的需要，构建了三个词典：情感词典、否定词词典、程度词词典。由于论坛语料的随意性较大以及金融领域情感分析的特殊性，现有的中文情感词典如 HowNet 和 NTUSD 难以满足金融情感分析的需要。为此，利用 SO-PMI[19]方法构建领域情感词典。SO-PMI 的思想是通过人工选取一组正向情感词(pLists)和一组负向情感词(nLists)作为基准词，根据待判定词语(word)与 pLists 和 nLists 之间的点间互信息差值，判定词语的情感倾向。以 N 表示语料库的文档总数，df(x&y)表示词 x 和 y 在语料库中共现的文档数，df(z)表示语料库中包含词 z 的文档数，计算公式如下：

$$\text{SO-PMI}(\text{word})=\sum_{p\in pLists}\log_2\frac{N\times df(\text{word}\,\&\,p)}{df(\text{word})\times df(p)}-\sum_{n\in nLists}\log_2\frac{N\times df(\text{word}\,\&\,n)}{df(\text{word})\times df(n)}$$

当某个词语的 SO-PMI 值大于 0 时，将其归为正向情感词，小于 0 则归为负向。通过对使用 SO-PMI 算法得到的情感词典进行人工筛选和调整，得到正向情感词 1 404 个，负向情感词 926 个。程度词词典下载自数据堂①，共 61 个。否定词词典通过人工添加得到，共 21 个。

3.2 依存句法

依存句法分析是自然语言处理中的一项重要技术，其任务是将输入的文本进行自动分析，得到文本的句法结构[20]。通过依存句法分析，可以了解句子中各词语之间的修饰关系，这种修饰关系可以非常方便地应用于句子情感倾向性分析。目前依存句法分析工具主要有哈尔滨工业大学的 LTP 语言云平台[21]、复旦大学 NLP 依存分析和 Stanford 句法分析器。其中，LTP 平台是国内较为成熟的中文自然语言处理平台，它提供了一套高效、准确、开放的文本处理模块，并在 SANCL 2012 互联网数据依存句法分析评测中取得第二名的成绩。从开放性和准确性考虑，本文选用 LTP 平台实现依存句法分析。LTP 中依存关系种类共有 14 种，如表 1 所示。

①http://www.datatang.com/data/44198.

表 1　LTP 依存句法标注关系

关系类型	标记	关系类型	标记
主谓关系	SBV	动补关系	CMP
动宾关系	VOB	并列关系	COO
间宾关系	IOB	介宾关系	POB
前置宾语	FOB	左附加关系	LAD
兼语	DBL	右附加关系	RAD
定中关系	ATT	独立结构	IS
状中关系	ADV	核心关系	HED

图 1 给出了一个 LTP 依存句法分析结果实例。每一个依存关系由核心词和修饰词组成，在 LTP 分析结果中，核心词由一条依存弧指向修饰词，依存弧上注明了具体的依存关系种类[22]。

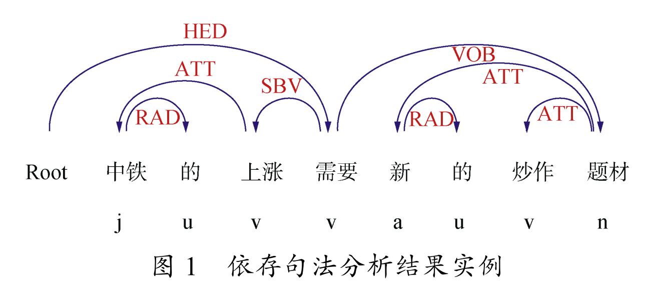


图 1　依存句法分析结果实例

3.3　情感主干抽取与情感传递

从汉语语法的角度看，句子中的主谓宾关系作为句子主干，基本表达了叙述者想表达的意思。以“不论最后如何发展，这对于大宗商品出口国显然是坏消息”为例。不考虑叙述者的针对对象，单纯只关心叙述者对发生事件的看法和态度，主谓宾关系“这是坏消息”能基本表达出叙述者的观点。然而，由于论坛短文本以及用语不规范的特性，一些句子不存在主谓宾关系。此时，根据 LTP 的分析结果，HED 关系是一很好的选择。根据 LTP 的输出，不管句子多么简略和不规范，都存在一个 HED(即句子核心)关系。HED 关系描述了整个句子的核心，概括了句子的中心思想，是了解叙述者态度的主要成分。基于上述考虑，本文提出如下情感主干抽取策略，即以句子的主谓宾关系为主干，若该句没有主谓宾，则以句子核心作为情感主干。

在获得情感主干之后，对每个主干词分别进一步提取其修饰词。通过情感传递将修饰词的情感值传递至该主干词。例如“上涨之门即将打开”，其主谓关系为“门打开”，本身不存在情感倾向，但“上涨”和主语“门”之间存在一条依存弧，即“上涨”是“门”的修饰词，可先将“上涨”的情感值传递给“门”，即计算组合“上涨+门”的情感值。此时，“门”具备了情感值，再看“门打开”时，它已带有情感。需注意的是，并不是所有依存关系都能传递情感。笔者参照万常选等[23]的研究，只考虑 6 种依存关系之间的情感传递，如表 2 所示:

表 2　影响文本情感倾向性的依存关系

关系类型	标记
主谓关系	SBV
动宾关系	VOB
动补关系	CMP
并列关系	COO
定中关系	ATT
状中关系	ADV

3.4　情感计算规则

事实上，在不同的依存关系以及不同的词性组合中，修饰词对被修饰的核心词(这里即为主干词)的情感影响是不相同的，也即存在情感传递差异。因此，需要结合依存关系的词性组合设定具体的情感计算规则。现有的研究大多单纯地从语言组合的角度分析依存关系的词性组合。然而论坛语料存在叙述随意、口语化严重的特征，单纯地从语言组合的角度难以概括依存关系的所有词性组合。为此，通过对大规模金融论坛语料的分析，笔者统计了可能影响文本情感倾向性的 6 种依存关系中出现的一些词性组合。表 3 列举了各种依存关系中词性组合频度最高的前 6 种。其中，词性组合的格式为“修饰词+核心词”。可以看出，除 ATT 关系外，其余依存关系的前 6 种高频词性组合累积频率均达 80%以上，而 ATT 关系也接近 60%。简单起见，仅对各依存关系的前 6 种高频词性组合设定情感计算规则，除此之外，一律以核心词和修饰词的情感分值相加作为该依存关系组合的情感分值。

在对大量金融论坛语料观察和分析的基础上，根据统计到的依存关系词性组合，同时借鉴文献[23]的相关研究成果，主要根据 6 类不同的词语间依存关系以及主干结构关系设定了 8 类情感计算规则。为叙述方便，以 S(·)表示词语或分句的情感分值，D(·)表示程度副词的程度值，P(·)表示词语的情感极性，Score 表示根据规则计算出的得分。另外，记修饰词为 mw，核心词为 cw，程度副词为 dd，否定副词为 nd，其他词性符号见表 3。

表 3　依存关系词性组合

依存关系	词性组合	说明	词性组合所占比例	依存关系	词性组合	说明	词性组合所占比例
ADV	d v	副词+动词	80.99%	SBV	n v	名词+动词	86.56%
	v v	动词+动词			r v	代词+动词	
	nt v	时间名词+动词			v v	动词+动词	
	p v	介词+动词			nh v	人名+动词	
	a v	形容词+动词			n a	名词+形容词	
	d a	副词+形容词			ns v	地理名称+动词	
ATT	n n	名词+名词	57.33%	VOB	n v	名词+动词	91.34%
	v n	动词+名词			v v	动词+动词	
	a n	形容词+名词			a v	形容词+动词	
	r n	代词+名词			r v	代词+动词	
	m n	数字+名词			m v	数字+动词	
	q n	量词+名词			q v	量词+动词	
COO	v v	动词+动词	90.96%	CMP	v v	动词+动词	93.44%
	n n	名词+名词			a v	形容词+动词	
	a a	形容词+形容词			p v	介词+动词	
	a v	形容词+动词			m v	数字+动词	
	j j	缩写+缩写			q v	量词+动词	
	nh nh	人名+人名			d v	副词+动词	

(1) ADV 类规则

ADV 为状中关系，修饰词作状语修饰核心词。当修饰词为副词时，副词使被修饰词的情感强度发生变化或极性反转。如“融资买入额太大，决定明天不进场”中的“太大”和“不进场”，程度副词“太”使情感词“大”的情感得到强化，因此词语组合的情感值可设为副词程度值和动词情感值的乘积。而否定副词“不”将“进场”的极性反转，因此其组合情感值可设为动词情感值的相反数。当词性组合为“形容词+动词”时，如“稳健接盘”、“成功突破”。由于形容词对动词具有一定的修饰作用，但重点仍在动词，因此将组合的情感值设为两者的加权求和，且形容词的权重低于动词，具体取值方法是前者为后者的一半。对“动词+动词”的组合，从所获得的语料来看，绝大多数情况两者的情感极性相同，且很难分辨谁更重要，如“进场抢筹”，因而以两者的和作为组合的情感值。而当修饰词为时间名词、介词、量词等时，由于这些修饰词一般不具有情感倾向，因此组合的情感值就等于核心词的情感值。因此，本类规则可表示如下：

if (mw, cw) is ((d, v) or (d, a)) and mw is dd
　then Score=D(mw)× S(cw);
if (mw, cw) is ((d, v) or (d, a)) and mw is nd　then Score= –S(cw);
if (mw, cw) is ((nt, v) or (p, v))　then Score = S(cw) ;
if (mw, cw) is (a, v)　then Score = 0.5 × S(mw) + S(cw) ;
if (mw, cw) is (v, v)　then Score = S(mw) + S(cw) ;

注意，本类规则中“动词+动词”这类修饰词与核心词词性相同的情况在其他下述的各类依存关系中也同样存在，而且大多都无法分辨两者情感谁更重要，因此后续都按相同方式进行处理。再有，修饰词为时间名词、介词、量词等情况时，其他下述各类规则中也存在类似情形，处理方式也一样，不赘述。

(2) ATT 类规则

ATT 为定中关系，是句中定语和中心语的关系。当修饰词为动词或形容词，核心词为名词时，动词或形容词作定语修饰名词。如“这是一个很大的阴谋”中的“大阴谋”，虽然“大”对“阴谋”具有修饰作用，但整个依存关系的情感倾向取决于名词“阴谋”的极性。因此，本类规则可表示如下：

if (mw, cw) is ((r, n) or (m, n) or (q, n))　then Score= S(cw);
if (mw, cw) is (n, n)　then Score = S(mw) + S(cw) ;
if (mw, cw) is ((v, n) or (a, n))　then Score = $|S(mw)|$ × P(cw) ;

(3) COO 类规则

COO 表示两个构成词语之间的平等关系。如“庄家故意压盘和打压”中的“压盘”与“打压”。因此以修饰

词和核心词的分值和作为组合情感值。

$$Score = S(mw) + S(cw)$$

(4) SBV 类规则

SBV 为主谓关系，是句子中的主干。当词性组合为“名词+动词”，“名词+形容词”时，整个依存关系的情感倾向很大程度上取决于名词的情感。如“多头不会轻易动摇”中的“多头”与“动摇”，“多头”的正向情感占有较大权重。因此，本类规则可表示如下：

if (mw, cw) is ((r, v) or (nh, v) or (ns, v)) then Score = S(cw);

if (mw, cw) is (v, v) then Score = S(mw) + S(cw);

if (mw, cw) is ((n, v) or (n, a)) then Score = S(mw) + 0.5 × S(cw);

(5) VOB 类规则

VOB 为动宾关系，当词性组合为“名词+动词”和“形容词+动词”时，名词或形容词为动词动作的承受对象，因而其情感主要体现在动词上。如“股价突破阻力”中的“突破”与“阻力”。因此，本类规则可表示如下：

if (mw, cw) is ((r, v) or (m, v) or (q, v)) then Score= S(cw);

if (mw, cw) is (v, v) then Score = S(mw) + S(cw);

if (mw, cw) is ((n, v) or (a, v)) then Score = 0.5×S(mw) + S(cw);

(6) CMP 类规则

CMP 为动补关系，即对动词所产生的动作进行补充说明。经过统计该关系在金融论坛语料中出现极少，所以，以修饰词和核心词的分值和作简单处理。

$$Score = S(mw) + S(cw)$$

(7) IS-DO 类规则

根据谓语的不同，主谓宾关系可以分为两大类：“是”类型和“做”类型。“是”类型(谓语为“是”、“就是”、“为了”等)是对主语是什么的解释说明，其重点在于宾语部分，如“利空是买入的绝佳机会”。而“做”类型则是对主语怎么样或在谓语动词的动作发生下做了什么的解释说明，如“主力正在拉升股价”。对于“是”类型的主谓宾关系，由于重点在宾语，所以在宾语为情感词的情况下，以宾语的情感分值作为整个关系的情感值，否则返回主语的情感分值。对于“做”类型的主谓宾关系，分别计算主语和谓语及谓语和宾语的情感分值，以两者之和作为最终的情感分值。

(8) 子句间规则

笔者发现，金融论坛语料较少出现转折等复杂的句式关系。为简化计算，规定整个句子的情感值为各个子句情感值之和。因而计算规则如下：

$$Score = S(s_1) + S(s_2) + \cdots + S(s_n)$$

其中，s_i 为各个子句。

3.5 情感计算模型

利用中国科学院计算技术研究所 NLPIR 汉语分词系统[24]对每条待分析文本进行分词，以 XML 的格式和 Post 的方式提交到 LTP 进行处理。根据 LTP 返回的结果，抽取句子情感主干，依照构建好的词典和情感计算规则，对情感主干进行词语间的情感传递及计算，最终得到整个句子的情感类别。本文的情感计算模型如图 2 所示：

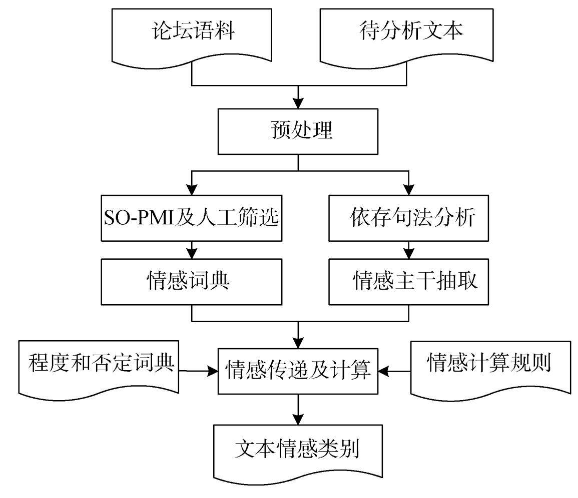


图 2　情感计算模型

4 实　验

4.1 实验语料

实验语料通过火车采集器采集自东方财富网股吧论坛①，选取生物医药板块的 5 只股票，以论坛中每个帖子的标题为采集对象，经过去噪后共得到 31 815 条数据，如表 4 所示：

表 4　5 家企业语料的数量

公司简称	数量
博雅生物	7 405
达安基因	8 603
国农科技	4 962
海王生物	5 334
华兰生物	5 511

①http: //guba.eastmoney.com/.

语料的标注由金融专业人士完成，分“强烈看涨”、“微弱看涨”、“中性”、“微弱看跌”、“强烈看跌”等 5 个情感级别进行标注。考虑到不同人士对中间三个级别的标注存在一定争议，而对“强烈看涨”和“强烈看跌”两个级别看法比较一致，实验仅选取各只股票数据中标注为“强烈看涨”和“强烈看跌”的语料并归为“看涨”(p)和“看跌”(n)两类，最后共得到 5 430 条数据，如表 5 所示：

表 5　实验数据分布

类别	数量
看涨	2 730
看跌	2 700
总计	5 430

4.2　实验结果及分析

文献[13]在挖掘股吧情绪时测试了 KNN、朴素贝叶斯、决策树、支持向量机 4 种常见的算法，结果表明 KNN 的准确率最高。为了验证本文方法的有效性，将其作为比较基准。此外，N-Gram 以统计词语间的依赖关系建立条件概率模型，也是一类常见的文本分类方法，且 Cui 等[25]认为 N>3 时能取得较好的效果，故笔者也将其作为比较基准。

实验时，每次将数据集的三分之二用于训练，三分之一用于测试。对每类方法，都采取随机抽样的方式，进行 10 次实验。记录每次实验测试集上的准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F 值(F-measure)等评测指标值。以本文方法的 10 次结果平均值与 KNN、N-Gram 准确率最高的一次进行比较。实验结果如表 6 所示：

表 6　实验结果对比

对比项目	本文方法	KNN	N-Gram
准确率(a)	84.46%	76.57%	71.04%
精确率(p)	82.84%	73.34%	71.83%
精确率(n)	86.28%	80.43%	70.30%
召回率(p)	87.14%	83.18%	69.78%
召回率(n)	81.74%	69.89%	72.33%
F 值(p)	84.94%	78.12%	70.79%
F 值(n)	83.95%	74.79%	71.30%

从表 6 可以看出，本文方法整体准确率为 84.46%，相较于 KNN 与 N-Gram 的文本分类方法，本文方法使得整体准确率有明显的提高。把看涨与看跌两类分开看，本文方法的看涨类召回率 87.14%，较 KNN 方法的看涨召回率 83.18%有较大的提升；而在看跌类召回率上，提升更为明显。F 值综合考虑了精确率和召回率，本文方法的看涨类 F 值为 84.94%，与 KNN(78.12%)和 N-Gram(70.79%)相比，分别提升了 6.82%和 14.15%。看跌类 F 值为 83.95%，与 KNN(74.79%)和 N-Gram(71.30%)相比，分别提升了 9.16%和 12.65%。这些结果都充分反映了基于句法结构信息的情感计算方法比起纯粹基于词频信息的机器学习方法有更好的优势。

5　结　语

本文基于依存句法，提出了一种针对金融论坛语料的情感分析方法。与机器学习方法相比较，在准确率、召回率和 F 值上均有较大提升，充分表明了句法结构与语义信息对文本情感分析的作用。

由于中文语言结构复杂，表达丰富多变，本文提出的方法对句法结构和语义关系信息仍没有充分挖掘。例如：如同文章和段落具有主题段和主题句一样，各子句对于整个句子的情感倾向性的贡献也不一样，本研究未区别对待；没有考虑主谓宾和句子核心的词语在情感传递后的词性改变问题；依赖于 LTP 的分析结果，虽然在现有各系统中其表现非常突出，但其准确性还有提升空间，相信随着其技术的进一步完善，可获得更好的结果。

未来研究将会重点关注金融论坛文本各子句对整个句子的情感权重以及设定更深层次的情感计算规则等问题。同时，情感分析技术的应用也是笔者的兴趣所在。

参考文献：

[1]　Liu B. Sentiment Analysis and Opinion Mining [M]. California: Morgan and Claypool Publishers, 2012.

[2]　Smailović J, Grčar M, Lavrač N, et al. Stream-based Active Learning for Sentiment Analysis in the Financial Domain [J]. Information Sciences, 2014, 285: 181-203.

[3]　Van de Kauter M, Breesch D, Hoste V. Fine-grained Analysis of Explicit and Implicit Sentiment in Financial News Articles [J]. Expert Systems with Applications, 2015, 42(11): 4999-5010.

[4] Hagenau M, Liebmann M, Neumann D. Automated News Reading: Stock Price Prediction Based on Financial News Using Context-capturing Features[J]. Decision Support Systems, 2013, 55(3): 685-697.

[5] 胡勇军, 江嘉欣, 常会友. 基于 LDA 高频词扩展的中文短文本分类[J]. 现代图书情报技术, 2013(6): 42-48. (Hu Yongjun, Jiang Jiaxin, Chang Huiyou. A New Method of Keywords Extraction for Chinese Short-text Classification [J]. New Technology of Library and Information Service, 2013(6): 42-48.)

[6] Kiritchenko S, Zhu X, Mohammad S M. Sentiment Analysis of Short Informal Texts [J]. Journal of Artificial Intelligence Research, 2014, 50: 723-762.

[7] 段江娇, 刘红忠, 曾剑平. 投资者情绪指数、分析师推荐指数与股指收益率的影响研究——基于我国东方财富网股吧论坛、新浪网分析师个股评级数据[J]. 上海金融, 2014(11): 60-64. (Duan Jiangjiao, Liu Hongzhong, Zeng Jianping. A Study of the Influence Between Investor Sentiment Index, Analyst Recommendation Index and Stock Index Return —— Based on Eastmoney.com and Sina Analyst Shares Rating Data[J]. Shanghai Finance, 2014 (11): 60-64.)

[8] 林炳灿. 基于投资者情绪的网络舆论对股票价格影响的统计研究[D]. 成都: 西南财经大学, 2013. (Lin Bingcan. Statistical Studies of Network Public Opinion's Investor Sentiment's Impact on the Stock Price[D]. Chengdu: Southwestern University of Finance and Economics, 2013.)

[9] 陈江鹏. 基于网络舆论的我国股票市场有效性检验研究[D]. 成都: 西南财经大学, 2013. (Chen Jiangpeng. A Study of China's Stock Market Effectiveness Testing Based on the Network Public Opinion [D]. Chengdu: Southwestern University of Finance and Economics, 2013.)

[10] 刘定平. 突发事件环境下投资者情绪对股票价格波动影响的实证研究[D]. 成都: 西南财经大学, 2014. (Liu Dingping. Empirical Research on the Volatility of Stock Price Brought by Investor's Sentiment in the Emergency Environment[D]. Chengdu: Southwestern University of Finance and Economics, 2014.)

[11] 张世军, 程国胜, 蔡吉花, 等. 基于网络舆情支持向量机的股票价格预测研究[J]. 数学的实践与认识, 2013, 43(24): 33-40. (Zhang Shijun, Cheng Guosheng, Cai Jihua, et al. Stock Price Prediction Base on Network Public Opinion and Support Vector Machine [J]. Mathematics in Practice and Theory, 2013, 43(24): 33-40.)

[12] 宋敏晶. 基于情感分析的股票预测模型研究[D]. 哈尔滨: 哈尔滨工业大学, 2013. (Song Minjing. Stock Prediction Model Based on Sentiment Analysis Research[D]. Harbin: Harbin Institute of Technology, 2013.)

[13] 金雪军, 祝宇, 杨晓兰. 网络媒体对股票市场的影响——以东方财富网股吧为例的实证研究[J]. 新闻与传播研究, 2013(12): 36-51. (Jin Xuejun, Zhu Yu, Yang Xiaolan. Effects of Online Media on Stock Market: An Empirical Study on Eastmoney.com[J]. Journalism & Communication, 2013(12): 36-51.)

[14] 沈翰彬. 投资者本地关注对股票收益率的影响——基于网络论坛文本挖掘的实证研究[D]. 杭州: 浙江大学, 2014. (Shen Hanbin. The Effect of Investor Home Attention on Stock Return – A Study Based on Stock Message Boards with Text Mining Techniques [D]. Hangzhou: Zhejiang University, 2014.)

[15] 夏梦南, 杜永萍, 左本欣. 基于依存分析与特征组合的微博情感分析[J]. 山东大学学报: 理学版, 2014, 49(11): 22-30. (Xia Mengnan, Du Yongping, Zuo Benxin. Micro-blog Opinion Analysis Based on Syntactic Dependency and Feature Combination [J]. Journal of Shandong University: Natural Science, 2014, 49(11): 22-30.)

[16] 张庆庆, 刘西林. 基于依存句法关系的文本情感分类研究[J]. 计算机工程与应用, 2015, 51(22): 28-32. (Zhang Qingqing, Liu Xilin. Sentiment Analysis Based on Dependency Syntactic Relation [J]. Computer Engineering and Applications, 2015, 51(22): 28-32.)

[17] Nakagawa T, Inui K, Kurohashi S. Dependency Tree-based Sentiment Classification Using CRFs with Hidden Variables [C]. In: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2010: 786-794.

[18] 肖红, 许少华. 基于句法分析和情感词典的网络舆情倾向性分析研究[J]. 小型微型计算机系统, 2014, 35(4): 811-813. (Xiao Hong, Xu Shaohua. Analysis on Web Public Opinion Orientation Based on Syntactic Parsing and Emotional Dictionary [J]. Journal of Chinese Computer Systems, 2014, 35(4): 811-813.)

[19] Turney P D, Littman M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[20] 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究[J]. 中文信息学报, 2007, 21(5): 73-79. (Yao Tianfang, Lou Decheng. Research on Semantic Orientation Analysis for Topics in Chinese Sentences [J]. Journal of Chinese Information Processing, 2007, 21(5): 73-79.)

[21] Che W X, Li Z H, Liu T. LTP: A Chinese Language Technology Platform [C]. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.2010: 13-16.

[22] 刘挺, 马金山. 汉语自动句法分析的理论与方法[J]. 当代语言学, 2009, 11(2): 100-112. (Liu Ting, Ma Jinshan. Theories and Methods of Chinese Automatic Syntactic Parsing: A Critical Survey[J]. Contemporary Linguistics, 2009, 11(2): 100-112.)

[23] 万常选, 江腾蛟, 钟敏娟, 等. 基于词性标注和依存句法的 Web 金融信息情感计算[J]. 计算机研究与发展, 2013, 50(12): 2554-2569. (Wan Changxuan, Jiang Tengjiao, Zhong Minjuan, et al. Sentiment Computing of Web Financial Information Based on the Part-of-Speech Tagging and Dependency Parsing[J]. Journal of Computer Research and Development, 2013, 50(12): 2554-2569.)

[24] NLPIR/ICTCLAS 汉语分词系统[EB/OL]. [2013-07-02]. http://ictclas.nlpir.org/. (NLPIR/ICTCLAS Chinese Segmentation System [EB/OL]. [2013-07-02]. http://ictclas.nlpir.org/.)

[25] Cui H, Mittal V, Datar M. Comparative Experiments on Sentiment Classification for Online Product Reviews[C]. In: Proceedings of the 21st National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2006: 1265-1270.

作者贡献声明:

兰秋军: 提出研究思路, 设计研究方案, 论文最终版本修订;
刘文星, 李卫康: 文献梳理;
刘文星, 李卫康, 胡星野: 采集、清洗和分析数据;
刘文星: 程序设计, 论文起草。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 兰秋军, 刘文星, 李卫康. dic.txt. 根据 SO-PMI 及人工处理产生的情感词典.

[2] 兰秋军, 刘文星, 李卫康. TestData.rar. 人工标注的测试数据.

[3] 兰秋军, 刘文星, 李卫康. 依存关系词性组合语料.rar. 分析词性组合频度的语料.

[4] 兰秋军, 刘文星, 李卫康. parsingCount.xls. 各依存关系中不同词性组合的频度.

收稿日期: 2015-10-14
收修改稿日期: 2016-01-10

Sentiment Analysis of Financial Forum Textual Message

Lan Qiujun Liu Wenxing Li Weikang Hu Xingye
(Business School, Hunan University, Changsha 410082, China)

Abstract: [**Objective**] This paper aims to identify sentiment propensity accurately with the help of a new method based on dependency parsing. [**Methods**] First, we extracted the sentiment stems of the sentences. Second, we defined sentiment-computing rules. Finally, we calculated sentiment propensity of each sentence. [**Results**] The proposed method achieved an overall accuracy of 84.46%. The average precision rate and recall rate for bullish class were 82.84% and 87.14% respectively, with an F-measure of 84.94%. In the mean time, bearish class got a precision rate of 86.28%, a recall rate of 81.74% and an F-measure of 83.95%. [**Limitations**] The proposed method did not consider the relevance among clauses. [**Conclusions**] The dependency parsing can effectively improve the accuracy of sentiment analysis of textual message from financial forum.

Keywords: Sentiment analysis Dependency parsing Financial forum text Text mining